Main Manuscript for

Tracing SARS-CoV-2 Clusters Across Local-scales Using Genomic Data

Leke Lyu[a], Guppy Stott[a], Cody Dailey[a], Sachin Subedi[a], Kayo Fujimoto[b], Ryker Penn[c], Pamela Brown[c], Roger Sealy[c], Justin Bahl[a]*

a. Center for Ecology of Infectious Diseases, Institute of Bioinformatics, Department of Infectious Diseases, Department of Epidemiology and Biostatistics, University of Georgia, Athens, GA, USA

b. Department of Health Promotion and Behavioral Sciences, The University of Texas Health Science Center at Houston, Houston, TX, USA

c. Houston Health Department, Houston, TX, USA

Corresponding Author: *Justin Bahl

Email: justin.bahl@uga.edu

Author Contributions: Leke Lyu and Justin Bahl conceptualized and designed research. Leke Lyu, Guppy Stott, Cody Dailey, Sachin Subedi and Kayo Fujimoto performed research. Ryker Penn , Pamela Brown and Roger Sealy contributed new data. Leke Lyu, Guppy Stott, Cody Dailey and Sachin Subedi wrote and reviewed the paper Justin Bahl acquired funding, supervised work, and coordinated communication among team members.

Competing Interest Statement: The authors declare that they have no conflict of interest.

Classification: Biological sciences / Ecology

Keywords: Viral evolution, Genomic epidemiology, Pandemic control

This PDF file includes:

Main Text
Figures 1 to 5

Abstract

Understanding local-scale transmission dynamics of SARS-CoV-2 is crucial for planning effective prevention strategies. This study analyzed over 26,000 genomes and their associated metadata collected between January and October 2021 to explore the introduction and dispersal patterns of SARS-CoV-2 in Greater Houston, a major metropolitan area noted for its demographic diversity. We identified more than a thousand independent introduction events, resulting in clusters of varying sizes, with earlier clusters presenting larger sizes and posing greater control challenges. Characterization of the sources of these introductions showed that domestic origins were more significant than international ones. Further examination of locally circulating clusters across different subregions of Greater Houston revealed varied transmission dynamics. Notably, subregions that served as primary viral sources sustained the local epidemic effectively, evidenced by: (1) a smaller proportion of new cases driven by external viral importations, and (2) longer persistence times of circulating lineages. Overall, our high-resolution spatiotemporal reconstruction of the epidemic in Greater Houston enhances understanding of the heterogeneous transmission landscape, providing key insights into regional response strategies and public health planning.

Significance Statement

The growing recognition of genome sequencing as critical for outbreak response has led to a rapid increase in the availability of sequence data. In this context, we put forward an analytical workflow within the Bayesian phylodynamic framework to identify and trace imported SARS-CoV-2 clusters using large-scale genome datasets. By utilizing metrics such as the Source-Sink Score, Local Import Score, and Persistent Time, our approach characterizes transmission patterns in each subregion and elucidates transmission heterogeneity. As new variants continue to emerge, the insights provided by our analysis are crucial for addressing the challenges of current and future pandemics effectively.

Main Text

Introduction

Genome epidemiology has significantly advanced our understanding and efforts to combat emerging infectious diseases (1–3). In the context of SARS-CoV-2, previous research has demonstrated its capability to clarify the virus's origins and spread (4, 5), reconstruct local transmission chains (6), assess the effectiveness of non-pharmaceutical interventions (7), and identify key predictors of viral lineage movements (8). These epidemiological insights, translated from the virus's evolutionary history, are crucial for shaping public health policies and would not have been possible without extensive sequencing efforts. By August 2024, over 16 million genome sequences had been submitted to the Global Initiative on Sharing All Influenza Data (GISAID) (9). Although these expansive COVID-19 datasets facilitate high-resolution inferences about local transmission dynamics, they also present significant computational challenges (10). In response, new algorithms, software, and computational workflows have emerged since the onset of the pandemic, including tools for rapid phylogenetic tree construction (11, 12) and the Thorney BEAST module for more efficient generation of time-resolved trees (13, 14).

Houston, the largest city in the Southern United States, anchors the Greater Houston metropolitan area and is one of the most demographically diverse cities in the country (15). It is also one of the most economically segregated cities, marked by sharp divisions in income, education, and occupation (16). According to Covid Act Now, Houston faces considerable challenges due to its high population density, significant proportion of non-English speakers, and notable income disparity. These factors strain the city's health system, making Houston more vulnerable to a SARS-

CoV-2 outbreak than over 90% of U.S. metropolitan areas (17). This vulnerability highlights the urgent need to understand local patterns of SARS-CoV-2 dispersal and how these patterns vary across different socioeconomic regions.

The COVID-19 pandemic has been characterized by the emergence and spread of genetically distinct virus variants that exhibit increased transmissibility compared to earlier lineages (18). The Delta variant, in particular, not only spreads more rapidly (19) but also leads to higher rates of hospitalization (20) and demonstrates greater immune evasion (21) than the previously dominant Alpha variant. In Houston, the emergence of Delta occurred in a context of heterogeneous prior immunity, resulting from both previous infections and vaccinations. The first case of the Delta variant in Houston Methodist Hospital was identified in mid-April 2021, during a period of declining COVID-19 cases (22). However, beginning in early July, there was a sharp increase in cases driven by the Delta variant, with these cases doubling in frequency approximately every seven days (22). Several critical questions remain unanswered, which cannot be resolved without genomic epidemiological inference: How long is the lag between a variant's introduction and its first clinical detection? What is the primary source of these variants? What role did Houston play in the introduction of Delta to the U.S.?

With the support of the Houston Health Department (HHD), we accessed an extensive dataset comprising over 10,000 Delta genomes sampled from Houston between January 2021 and October 2021, each linked with metadata such as zip code, age, and sex. This dataset provides a valuable opportunity to investigate transmission dynamics in Houston. It enables us to examine how population structure influences disease spread and to assess variations in SARS-CoV-2 transmission across different subregions.

A notable phylodynamic workflow developed by Simon Dellicour (23) facilitates large-scale phylogeographic analysis through two principal steps: first, a preliminary discrete trait analysis (24) on fixed empirical topologies identifies introduction events; second, it estimates the circulation dynamics of local viral clusters (25, 26). We adapted this analytical workflow to examine the spatial invasion dynamics of SARS-CoV-2 in Greater Houston, as illustrated in Figure 1. Utilizing viral genetic sequence data isolated from patients, our study aimed to determine the timing and number of viral introductions during the outbreak. We specifically investigated whether international or domestic importation played a more significant role. Additionally, we modeled the transmission structure among different demographic groups, including sex and age. Finally, we explored the spatiotemporal variation of viral dispersal across various subregions, such as Independent School Districts or counties, in Greater Houston.

Results

Detect Distinct Introduction Events and Identify Locally Circulating Clusters

Our dataset comprised 26,138 SARS-CoV-2 complete genomes, including 9,186 sampled from Houston and 16,952 contextual sequences. We conducted a discrete phylogeographic analysis on the time-calibrated phylogeny to identify distinct SARS-CoV-2 introduction events into Houston. This analysis revealed a total of 1,479 independent introduction events (95% highest posterior density [HPD]: 1,402 to 1,556). Notably, the sizes of resulting circulating clusters were highly skewed (Figure S1). The majority of introductions (909 events, 95% HPD: 853 to 968) resulted in singletons, while a few introductions led to clusters exceeding 2,000 cases. Temporal analysis further revealed that earlier introductions were more likely to result in larger clusters (Figure 2A). Specifically, during EPI Week 17, 55.6% of introductions resulted in clusters larger than 10, 33.3% in clusters smaller than 10, and 11.1% were singletons. By EPI Week 30, the distribution had shifted

significantly, with 86.2% resulting in singletons, 13.8% forming clusters smaller than 10, and no clusters larger than 10.

We classified viral imports into Houston based on their origin: domestic or international. At the onset of the outbreak, we observed scattered introductions from both sources. After late April, domestic importations rapidly increased and became the dominant type (Figure 2B). There were a total of 1,359 domestic imports (95% HPD: 1,279 to 1,432), significantly outnumbering the 119 international imports (95% HPD: 109 to 132). However, international imports peaked earlier and were associated with larger mean cluster sizes (Figures S2 and S3).

Subtrees extending from the introduction nodes were identified as locally circulating clusters. We selected 181 clusters, each containing five or more isolates, for further analysis of local dispersal. In total, these separate clusters included 7,657 sequences from the Greater Houston area, with the two largest clusters containing 2,198 and 2,031 sequences, respectively.

Phylogeny-Trait Correlation Among Locally Circulating Clusters

We explored the correlations between phylogenetic structures and demographic traits—specifically age and sex—to enhance our understanding of the factors influencing transmission dynamics. Our null hypothesis is that these traits are randomly associated with the phylogenetic structures. A low p-value ($p < 0.05$) refutes this hypothesis, indicating a strong correlation and suggesting limited viral dispersal between different traits.

We tested the association between 181 locally circulating clusters and demographic traits (Figure 3). In 20 clusters, the age group traits were tightly correlated with the phylogeny ($p < 0.05$). In contrast, only 6 clusters showed a similar tight correlation for sex traits. Generally, sex groups appear more interspersed within the phylogeny, suggesting that viral transmission is more constrained within age groups than between sex groups.

Demographic Determinants of Transmission

Locally circulating clusters encompassed 7,657 genomes distributed among diverse age groups: 2,412 young adults (ages 19–35), 2,342 middle-aged adults (ages 36–55), 995 infants and children (ages 0–12), 963 seniors (ages 56 and over), 919 teenagers (ages 13–18), and 26 individuals of unknown age. The sex distribution included 3,954 males, 3,687 females, and 16 individuals of unknown sex.

We jointly estimated a single discrete trait model to all circulating clusters to quantify viral dispersal among age and sex groups. We found young and middle-aged adults were identified as the primary drivers of viral transmission (Figure 4A). The eight most significant transitions included: from young to middle-aged adults at a rate of 4.775 transitions per year, and vice versa at 2.194 transitions per year; from middle-aged adults to infants and children at 1.946 transitions per year; to seniors at 1.567 transitions per year; to teenagers at 1.494 transitions per year; from young adults to seniors at 1.363 transitions per year; to teenagers at 1.245 transitions per year; and to infants and children

at 1.174 transitions per year. Each of these transitions was strongly supported by a Bayes Factor exceeding 100. The detailed list for diffusion rate among age groups are available is Table S1.

In the sex-based model (Figure 4B), we observed that the transition rate from females to males (1.652) was higher than from males to females (0.437). The detailed list for diffusion rate among sex groups are available is Table S2.

Heterogeneous Dynamics of Viral Dispersal in Subregions of Greater Houston

We estimated a jointly fitted single discrete model to reconstruct the viral dispersal history in Greater Houston. This model categorized location traits into 29 groups, including 21 Independent School Districts in Harris County and 8 nearby counties. The joint model (Figure 5A) for these subregions revealed 82 transitions that were decisively supported by Bayes factors (>100). The transition from Fort Bend County to the Houston ISD had the highest transition rate, at 23.780. This was followed by transitions from Houston ISD to Cypress-Fairbanks ISD, with a rate of 10.203, and from Montgomery County to Houston ISD, with a rate of 6.923. The detailed list of rates can be found in Table S3.

We calculated the Source-Sink Scores (SSS) to identify populations as either viral sources or sinks, based on the net viral flow weighted by outbreak size. Using this metric, we ranked subregions from the most dominant sources to sinks (Figure 5B). We identified Houston ISD, with a Source-Sink Score of 0.629 (95% HPD: 0.549 to 0.708), Fort Bend County, with a score of 0.550 (95% HPD: 0.381 to 0.651), and Cypress-Fairbanks ISD, with a score of 0.069 (95% HPD: -0.020 to 0.171), as the primary sources for local dispersal in the Greater Houston area, where Source-Sink Scores were greater than 0. Further, we calculated Local Import Scores (LIS) to assess the relative influence of viral introductions versus local transmission in driving the epidemic (Figure 5C). Viral sources exhibited the lowest Local Import Scores, indicative of strong, locally sustained transmission. In contrast, viral sinks with higher Local Import Scores relied more on external introductions. Additionally, we estimated the median Persistence Time (PT) of viral transmission chains within each subregion (Figure 5D). The primary viral sources demonstrated the longest persistence times, showing more successful local transmission.

Discussion

The global spread of SARS-CoV-2 triggers new outbreaks, but the majority of cases result from local transmission. Quantitatively understanding the local transmission dynamics is crucial for informing effective prevention.

In collaboration with the Houston Health Department, we analyzed over 26,000 genomes and their associated metadata to study the introduction and dispersal of SARS-CoV-2 in Greater Houston—a major metropolitan area known for its demographic diversity—between January and October 2021. Our analysis identified 1,479 independent introduction events (95% HPD: 1,402 to 1,556). Characterizing the sources of these introductions revealed that domestic origins were the predominant source overall (Figure 2). However, international importations led to more successful local transmission, as evidenced by larger cluster sizes. We also assessed how demographic structures influence the dynamics of disease spread. The tip-trait association test suggests that viral transmission is more restricted within age groups than between sexes (Figure 3). Additionally, the discrete trait analysis modeled transmission between different demographic categories (Figure 4). Finally, we reconstructed the spatiotemporal dispersal of pre-identified local outbreak clusters (Figure 5), revealing heterogeneous transmission dynamics across subregions of Greater Houston. Specifically, in Houston ISD, Fort Bend County, and Cypress-Fairbanks ISD - identified as key viral sources - introductions accounted for a smaller percentage of new cases and exhibited longer chains of local transmission.

Our analysis quantitatively confirms that the Delta outbreak in Houston was driven by multiple independent introduction events. These importations led to widespread local transmission, resulting in clusters of varying sizes, with earlier clusters being larger and more difficult to eliminate. This pattern aligns with previous findings on COVID-19 introductions in the UK (8, 27), New York City (28), and Denmark (5). At the onset of the outbreak, introductions came from both domestic and international sources. After late April, however, domestic importations surged and became dominant. Since we observed no clear predominance of international sources throughout the outbreak, we believe that Houston likely did not serve as a primary entry point into the U.S., unlike New York, California, and Florida, which have been identified as major entry points in previous research (4). Nevertheless, we cannot overlook the impact of international importation, particularly in the early stages, as it generally led to larger clusters and more sustained local transmission, placing a significant burden on public health intervention.

Tip-trait association quantifies the degree to which viral phenotypic characters are correlated with shared ancestry, as represented by a viral phylogenetic tree (29). A common application of this phylogeny-trait correlation is to explore spatial structure (30); specifically, whether sequences group together in a phylogeny based on geographic location. In this study, we examine the correlations between phylogeny and population structures to better understand how demographic factors influence transmission dynamics. Human movements and interactions were generally more constrained by age group than by sex. For example, children are typically found in daycare centers or middle schools, teenagers in high schools, adults at their workplaces, and seniors in nursing homes. Our analysis on circulating clusters statistically supports that traits age groups are more tightly correlated with the tree topology, indicating more constrained transmission within these groups. Furthermore, we estimated the discrete trait model to describe the transmission between age and sex groups. We found that young and middle-aged adults were identified as the primary drivers of viral transmission. SARS-CoV-2 disproportionately affects men more than women (31), and our findings revealed that the transition rate from females to males was higher than from males to females. This aligns with medical observations that males are more susceptible than females (32).

Previous report showed the COVID-19 epidemic in Houston exhibits distinct patterns, including varying infection probabilities and hospitalization rates (22, 33). Here, we reconstructed spatial dispersal of SARS-CoV-2 across 29 subregions of Greater Houston using discrete trait analysis, applying the Source-Sink Score, Local Import Score, and Persistent Time to characterize transmission patterns in each subregion. These metrics, integrated with Bayesian phylogeographic inference, allowed us to calculate their values along with their highest posterior density intervals, providing a measure of confidence in our estimates. Our analysis revealed a consistent pattern across all subregions: regions with higher Source-Sink Scores were associated with lower Local Import Scores and higher Persistent Times. This pattern aligns with previous analyses of viral transmission in Seattle (34), where a structured coalescent model (35) found that South King County exhibited longer persistence of local transmission compared to North King County, where external viral importations drove a larger proportion of new cases. As the Source-Sink Score provides a heuristic understanding for identifying whether a region functions as either a source or sink of viral transmission, these findings collectively suggest that a well-established and sustained local epidemic is crucial for a region to act as a source of pathogen spread to other areas. Targeted public health interventions in these identified source regions—such as temporary closures of schools, limitations on large public gatherings, enhanced testing and contact tracing, and increased access to healthcare resources—could not only mitigate local transmission but also have a broader impact by reducing the spread of the virus across the entire Greater Houston area. Such focused strategies can enhance the efficiency of outbreak control measures and allocate resources more effectively to areas with the greatest influence on regional transmission dynamics.

Materials and Methods

SARS-CoV-2 Genomic Dataset

With the support of the Houston Health Department (HHD), we accessed a large dataset of SARS-CoV-2 genomes sampled in Houston (>10,000), along with linked metadata, including ZIP code, age group, and sex. The first reported Delta variant case in Houston occurred in mid-April 2021. Our contextual dataset (non-Houston) was divided into two phases: Phase one included all worldwide sequences available in GISAID (www.gisaid.org) sampled before April 15, while Phase two sampled 1% of worldwide sequences available after April 15 (Figure S4). This sampling scheme balanced the need to include early sequences with the practical limits of handling a rapidly growing dataset.

The combined dataset of Delta sequences from Houston and contextual sources was then aligned to the reference genome (GenBank ID: NC_045512.2) using minimap2 v2.24 (36). We filtered out low-quality sequences with mapped completeness below 93% and trimmed alignments outside the reference coordinates 265:29674, padding with Ns to mask out the 3' and 5' UTRs. We then calculated the genetic base-pair differences between the alignments and the reference genome. For samples collected within the same Epi-Week, we excluded sequences with genetic differences greater than 3.0 standard deviations from the mean, aiming to preliminarily filter out those with a poor clock signal. In total, 26,138 alignments passed the filtering criteria. The contextual sequences were categorized as either domestic (excluding Houston) or international. Our dataset included 9,186 sequences from Houston, 5,334 from domestic sources, and 11,618 from international sources (Figure S5).

Time-Scaled Phylogenetic Inference

We inferred the time tree of our dataset in two steps. First, a maximum-likelihood phylogeny was estimated using IQ-TREE 2.3.2 (37) with the default settings. Subsequently, the resulting tree was time-calibrated using BEAST v1.10.5 (38). Given the size of our dataset, Thorney BEAST (13, 14) was applied to significantly reduce computational time by employing an alternative likelihood function. XML files for the BEAST runs were prepared using R scripts. Phylogeny file editing were performed using the tools jclusterfunk v0.0.25 (39) and gotree v0.4.5 (40). To minimize runtime by reducing burn-in, a roughly scaled time tree estimated by TreeTime 0.11.2 (41) was included in the XML file as the starting tree. We executed five chains of 2.5 billion states each, with a burn-in of 1 billion states. Trees were sampled every 7.5 million states, resulting in an empirical tree set size of 1,000. The convergence and mixing of all relevant parameters were inspected using Tracer 1.7 (42) to ensure that their associated effective sample size (ESS) values exceeded 200. One posterior tree (Figure S6) was then randomly selected and used as a fixed time tree for subsequent introduction analysis.

Introduction Analysis

We performed a preliminary discrete phylogeographic analysis on the fixed time tree obtained in the previous step to identify descent clusters likely arose from independent introductions. Tips of the tree were assigned as either 'Houston', 'Domestic', or 'International'. XML files for BEAST runs were generated using R scripts. We executed five chains of 20 million states each, with a burn-in of 4 million states. Trees then were sampled every 160,000 states, resulting in an empirical tree set size of 500. The convergence and mixing aspects of all relevant parameters were inspected using Tracer 1.7 (42).

We considered an introduction event when a node was assigned the location 'Houston', while its parent node was labeled as non-Houston (either 'Domestic' or 'International'). We defined locally circulating clusters as the subtree extending downward from the introduction node. Introduction

time was identified as the midpoint on the branch connecting the introduction node and its parent. We summarized the weekly count of introduction events from the posterior tree sets, providing 95% highest posterior density estimates for these counts. Phylogeny reading and manipulation, including summarizing cluster size, were performed using ape 5.8 (43), treeio 1.20.2 (44) and ggtree 1.14.6 (45) packages.

We selected a representative tree (Figure S7) from the posterior tree set that matched the posterior median for total importations. From this tree, we extracted 181 distinct locally circulating clusters with five or more sequences. In each cluster, tips sampled outside Houston were pruned using the ape 5.8 package (43).

Tip-Trait Association Test

For each locally circulating cluster, we used the Association Index (29, 46) to quantify phylogeny-trait correlations. To assess the significance of these correlations, we generated null distributions of the Association Index by randomizing trait assignments on the tips 1,000 times, enabling us to perform the tip-trait association test. All scripts for this test were bundled into an R package named TTAT, which is publicly available on GitHub at https://github.com/leke-lyu/TTAT. This package takes various clades and trait data as input, using metrics such as the Association Index and parsimony score to perform the test. The p-values from the tests are displayed in a scatter plot of sex group vs. age group using ggplot2 (47).

Jointly Fitted Discrete Trait Model

Discrete trait analysis models the evolution of discrete states on a phylogeny, similar to sequence evolution. Assuming all circulating clusters shared underlying characteristics, we jointly estimated a single transition rate matrix (28, 48, 49) to describe transitions between traits—location, sex, and age groups—with the aim of reconstructing geographic dispersal patterns and identifying demographic determinants of transmission. XML files for BEAST runs were created using R scripts. For each trait, we ran five chains of 100 million states each, with a burn-in period of 20 million states. Trees were subsequently sampled every 800,000 states, ensuring each cluster achieved an empirical tree set size of 500. Convergence and mixing of all relevant parameters were inspected using Tracer.

Over half of the Greater Houston population resides in Harris County (50). In our phylogeographic model, we divided Harris County into 21 Independent School Districts (ISDs) and combined these with the eight surrounding counties, resulting in a total of 29 trait categories. When translating ZIP codes (the sampling location record) into these subregions, some ZIP codes spanned multiple ISDs; for instance, parts of one ZIP code might fall in both ISD M and ISD N. We treated these intersections as ambiguous traits, allowing the model to interpret the trait as either M or N. Additionally, 16 sequences lacked associated sex records, and 26 sequences lacked age data, which we also categorized as ambiguous traits.

Posterior Processing

The joint estimation procedure reconstructed the ancestral states for 181 clusters. Given these posterior tree sets, we estimate the following epidemiological metrics:

A. Source Sink Score (51)**:** This metric, ranging between -1 and 1, measures net viral exports, weighted by outbreak size. A score approaching 1 suggests that the region primarily functions as a viral source. Conversely, a score nearing -1 indicates that the region predominantly acts as a viral sink.

B. Local Import Score (51)**:** Ranging from 0 to 1, this metric measures the fraction of introductions relative to the total count of new cases in a region. A score near 1 indicates that importations predominate, while a score near 0 suggests that local transmissions dominate, indicating that the epidemic is primarily sustained locally.

C. Persistence Time: This metric measures how long a viral lineage circulates in a region. It is calculated by tracing the number of days it takes for a tip to move from its sampled location, going backward up the phylogeny until the node location differs from the tip location (34, 52).

Given the size of our dataset, we employed a divide and conquer strategy to efficiently manage the input data. The empirical tree sets for each local cluster were divided into separate files, with each file containing a tree corresponding to a unique state. Tree files shared the same state were read into 'phylo' objects using the treeio 1.20.2 package (44). These 'phylo' objects were then converted into structured data frames using the tidytree 0.4.6 package (53), facilitating the easier estimation of epidemiological metrics.

Acknowledgments

This work has been funded in part from the National Institute of Allergy and Infectious Diseases, a component of the NIH, Department of Health and Human Services, under contract no. 75N93021C00018 (NIAID Centers of Excellence for Influenza Research and Response, CEIRR) and Centers for Disease Control and Prevention, Department of Health and Human Services, under contracts 75D30121C10133 and NU50CK000626. We acknowledge the GISAID contributors (acknowledgment table of genomes used is provided on our GitHub repository) for sharing genomic data.

References

1. S. W. Attwood, S. C. Hill, D. M. Aanensen, T. R. Connor, O. G. Pybus, Phylogenetic and phylodynamic approaches to understanding and combating the early SARS-CoV-2 pandemic. *Nat. Rev. Genet.* **23**, 547–562 (2022).

2. V. Hill, C. Ruis, S. Bajaj, O. G. Pybus, M. U. G. Kraemer, Progress and challenges in virus genomic epidemiology. *Trends Parasitol.* **37**, 1038–1049 (2021).

3. N. D. Grubaugh, *et al.*, Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10–19 (2019).

4. T. Alpert, *et al.*, Early introductions and transmission of SARS-CoV-2 variant B.1.1.7 in the United States. *Cell* **184**, 2595-2604.e13 (2021).

5. T. Y. Michaelsen, *et al.*, Introduction and transmission of SARS-CoV-2 lineage B.1.1.7, Alpha variant, in Denmark. *Genome Med.* **14**, 47 (2022).

6. P. Lemey, *et al.*, Untangling introductions and persistence in COVID-19 resurgence in Europe. *Nature* **595**, 713–717 (2021).

7. N. F. Müller, *et al.*, Viral genomes reveal patterns of the SARS-CoV-2 outbreak in Washington State. *Sci. Transl. Med.* **13**, eabf0202 (2021).

8. J. L.-H. Tsui, *et al.*, Genomic assessment of invasion dynamics of SARS-CoV-2 Omicron BA.1. *Science* **381**, 336–343 (2023).

9. S. Elbe, G. Buckland-Merrett, Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46 (2017).

10. B. Morel, *et al.*, Phylogenetic Analysis of SARS-CoV-2 Data Is Difficult. *Mol. Biol. Evol.* **38**, 1777–1791 (2020).

11. Y. Turakhia, *et al.*, Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. *Nat. Genet.* **53**, 809–816 (2021).

12. N. De Maio, *et al.*, Maximum likelihood pandemic-scale phylogenetics. *Nat. Genet.* **55**, 746–752 (2023).

13. X. Didelot, I. Siveroni, E. M. Volz, Additive Uncorrelated Relaxed Clock Models for the Dating of Genomic Epidemiology Phylogenies. *Mol. Biol. Evol.* **38**, 307–317 (2021).

14. X. Didelot, N. J. Croucher, S. D. Bentley, S. R. Harris, D. J. Wilson, Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).

15. X, Instagram, Email, Facebook, How Houston has become the most diverse place in America. *Los Angel. Times* (2017). Available at: https://www.latimes.com/nation/la-na-houston-diversity-2017-htmlstory.html [Accessed 5 August 2024].

16. C. Wilson, These Are the Most Economically Segregated Cities in America. *TIME* (2017). Available at: https://time.com/4744296/economic-segregation-cities-america/ [Accessed 5 August 2024].

17. U.S. COVID Risk & Vaccine Tracker. *Covid Act Now*. Available at: https://covidactnow.org [Accessed 5 August 2024].

18. J. T. McCrone, *et al.*, Context-specific emergence and growth of the SARS-CoV-2 Delta variant. *Nature* **610**, 154–160 (2022).

19. P. Elliott, *et al.*, Exponential growth, high prevalence of SARS-CoV-2, and vaccine effectiveness associated with the Delta variant. *Science* **374**, eabl9551 (2021).

20. K. A. Twohig, *et al.*, Hospital admission and emergency care attendance risk for SARS-CoV-2 delta (B.1.617.2) compared with alpha (B.1.1.7) variants of concern: a cohort study. *Lancet Infect. Dis.* **22**, 35–42 (2022).

21. C. Lucas, *et al.*, Impact of circulating SARS-CoV-2 variants on mRNA vaccine-induced immunity. *Nature* **600**, 523–529 (2021).

22. P. A. Christensen, *et al.*, Delta Variants of SARS-CoV-2 Cause Significantly Increased Vaccine Breakthrough COVID-19 Cases in Houston, Texas. *Am. J. Pathol.* **192**, 320–331 (2022).

23. S. Dellicour, *et al.*, A Phylodynamic Workflow to Rapidly Gain Insights into the Dispersal History and Dynamics of SARS-CoV-2 Lineages. *Mol. Biol. Evol.* **38**, 1608–1613 (2021).

24. P. Lemey, A. Rambaut, A. J. Drummond, M. A. Suchard, Bayesian Phylogeography Finds Its Roots. *PLOS Comput. Biol.* **5**, e1000520 (2009).

25. O. G. Pybus, *et al.*, Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci.* **109**, 15066–15071 (2012).

26. P. Lemey, A. Rambaut, J. J. Welch, M. A. Suchard, Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877–1885 (2010).

27. L. du Plessis, *et al.*, Establishment and lineage dynamics of the SARS-CoV-2 epidemic in the UK. *Science* **371**, 708–712 (2021).

28. S. Dellicour, *et al.*, Variant-specific introduction and dispersal dynamics of SARS-CoV-2 in New York City – from Alpha to Omicron. *PLOS Pathog.* **19**, e1011348 (2023).

29. J. Parker, A. Rambaut, O. G. Pybus, Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infect. Genet. Evol.* **8**, 239–246 (2008).

30. E. C. Holmes, The phylogeography of human viruses. *Mol. Ecol.* **13**, 745–756 (2004).

31. I. Lakbar, D. Luque-Paz, J.-L. Mege, S. Einav, M. Leone, COVID-19 gender susceptibility and outcomes: A systematic review. *PLoS ONE* **15**, e0241827 (2020).

32. C. Russo, *et al.*, Candidate genes of SARS-CoV-2 gender susceptibility. *Sci. Rep.* **11**, 21968 (2021).

33. Harris County COVID-19 Data Hub 1. Available at: https://covid-harriscounty.hub.arcgis.com/ [Accessed 27 July 2024].

34. M. I. Paredes, *et al.*, Local-scale phylodynamics reveal differential community impact of SARS-CoV-2 in a metropolitan US county. *PLOS Pathog.* **20**, e1012117 (2024).

35. N. F. Müller, D. Rasmussen, T. Stadler, MASCOT: parameter and state inference under the marginal structured coalescent approximation. *Bioinformatics* **34**, 3843–3848 (2018).

36. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

37. L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, B. Q. Minh, IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

38. M. A. Suchard, *et al.*, Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016 (2018).

39. snake-flu/jclusterfunk. Available at: https://github.com/snake-flu/jclusterfunk [Accessed 28 July 2024].

40. Gotree/Goalign: toolkit and Go API to facilitate the development of phylogenetic workflows | NAR Genomics and Bioinformatics | Oxford Academic. Available at: https://academic.oup.com/nargab/article/3/3/lqab075/6348148 [Accessed 28 July 2024].

41. P. Sagulenko, V. Puller, R. A. Neher, TreeTime: Maximum-likelihood phylodynamic analysis. *Virus Evol.* **4**, vex042 (2018).

42. A. Rambaut, A. J. Drummond, D. Xie, G. Baele, M. A. Suchard, Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).

43. E. Paradis, K. Schliep, ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).

44. L.-G. Wang, *et al.*, Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol. Biol. Evol.* **37**, 599–603 (2020).

45. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

46. T. H. Wang, Y. K. Donaldson, R. P. Brettle, J. E. Bell, P. Simmonds, Identification of Shared Populations of Human Immunodeficiency Virus Type 1 Infecting Microglia and Tissue Macrophages outside the Central Nervous System. *J. Virol.* **75**, 11686–11699 (2001).

47. H. Wickham, *ggplot2* (Springer International Publishing, 2016).

48. K. Fujimoto, *et al.*, Methodological synthesis of Bayesian phylodynamics, HIV-TRACE, and GEE: HIV-1 transmission epidemiology in a racially/ethnically diverse Southern U.S. context. *Sci. Rep.* **11**, 3325 (2021).

49. J. Bahl, *et al.*, Influenza A Virus Migration and Persistence in North American Wild Birds. *PLOS Pathog.* **9**, e1003570 (2013).

50. Harris County, Texas - Census Bureau Profile. Available at: https://data.census.gov/profile/Harris_County,_Texas?g=050XX00US48201 [Accessed 30 July 2024].

51. L. Lyu, *et al.*, Characterizing SARS-CoV-2 Transmission Heterogeneity Between Urban and Rural Populations in Texas, USA, Using a Novel Spatial Transmission Count Statistic. [Preprint] (2024). Available at: https://www.medrxiv.org/content/10.1101/2023.12.28.23300535v3 [Accessed 15 July 2024].

52. T. Bedford, S. Cobey, P. Beerli, M. Pascual, Global Migration Dynamics Underlie Evolution and Persistence of Human Influenza A (H3N2). *PLOS Pathog.* **6**, e1000918 (2010).

53. G. Yu, *Data Integration, Manipulation and Visualization of Phylogenetic Trees* (Chapman and Hall/CRC, 2022).

Figures and Tables

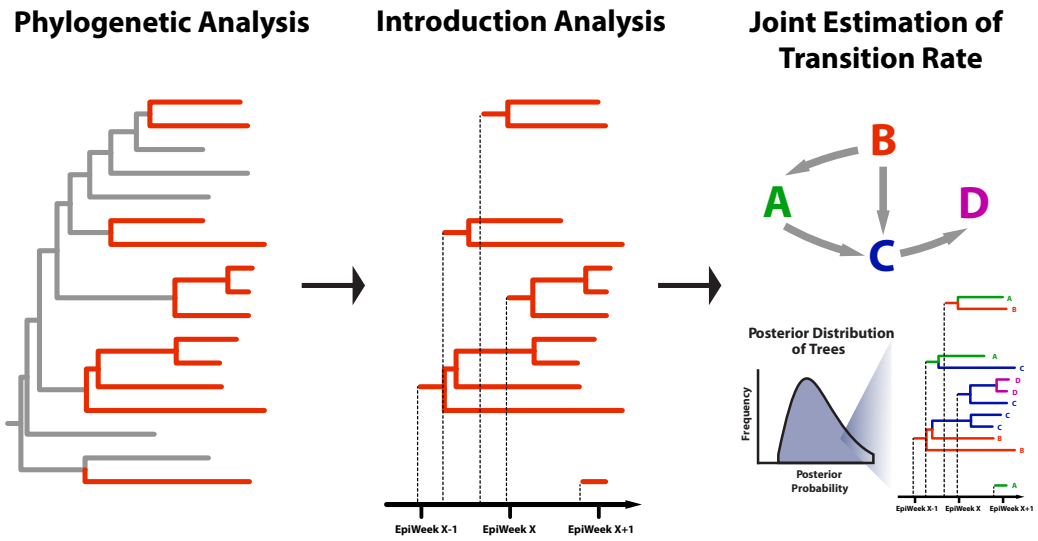

Figure 1. Conceptual Workflow of Large-Scale Genomic Epidemiology Analysis. Phylogenetic Analysis built the phylogeny of the isolated sampled from the focal region within a global context. **Introduction analysis** estimated the timing of viral introductions and identified locally circulating clusters. **The Joint Fit Model** inferred transition rate under a unified transition matrix.

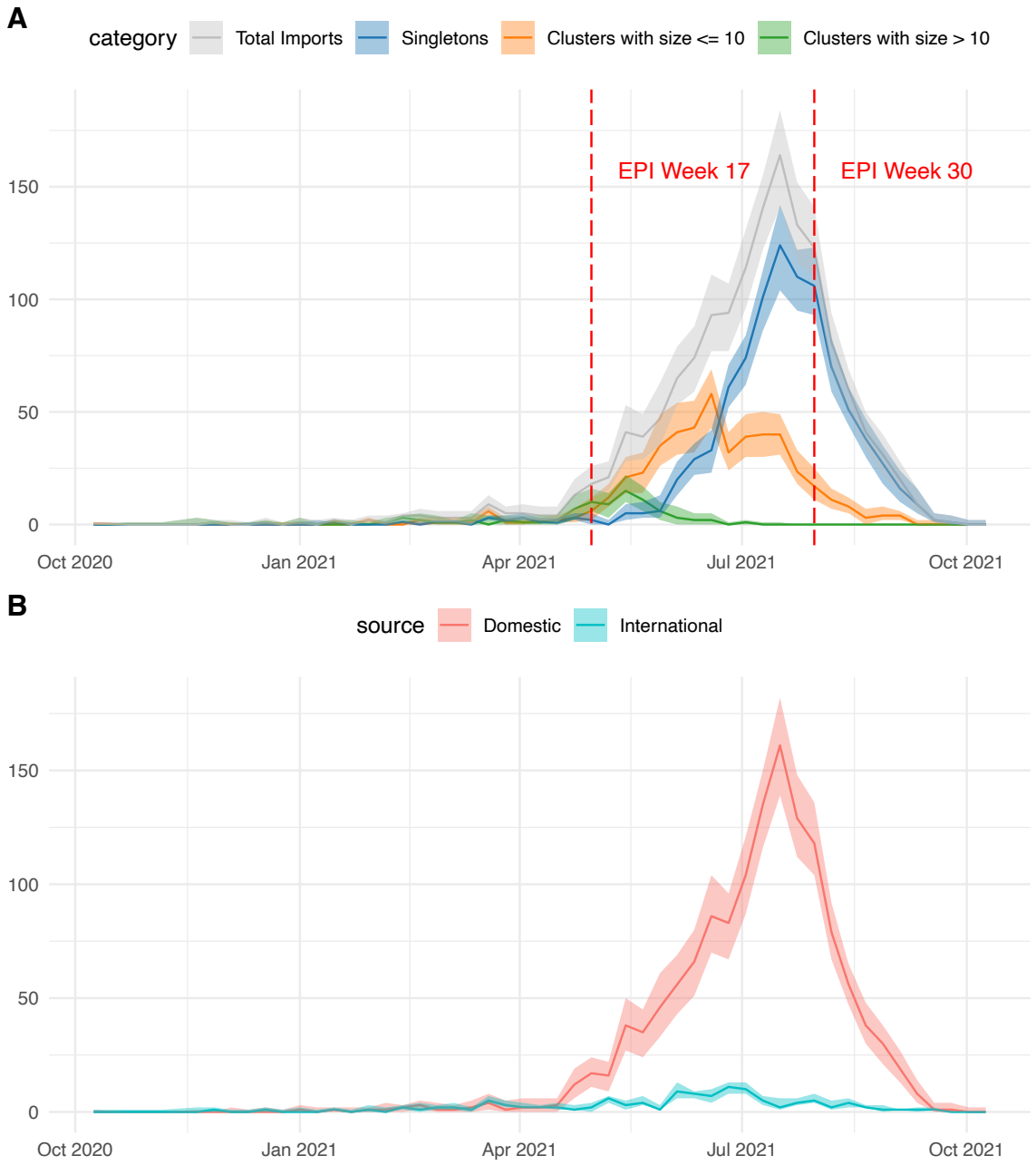

Figure 2. Dynamics of Viral Introduction into the Greater Houston. A. Estimated weekly frequency of viral introductions, with curves colored by the size of the resulting local transmission lineages. Shading indicates the associated 95% Highest Posterior Density (HPD). **B.** Estimated weekly frequency of viral introductions, with curves colored to distinguish between domestic and international sources. Shading indicates the associated 95% HPD.

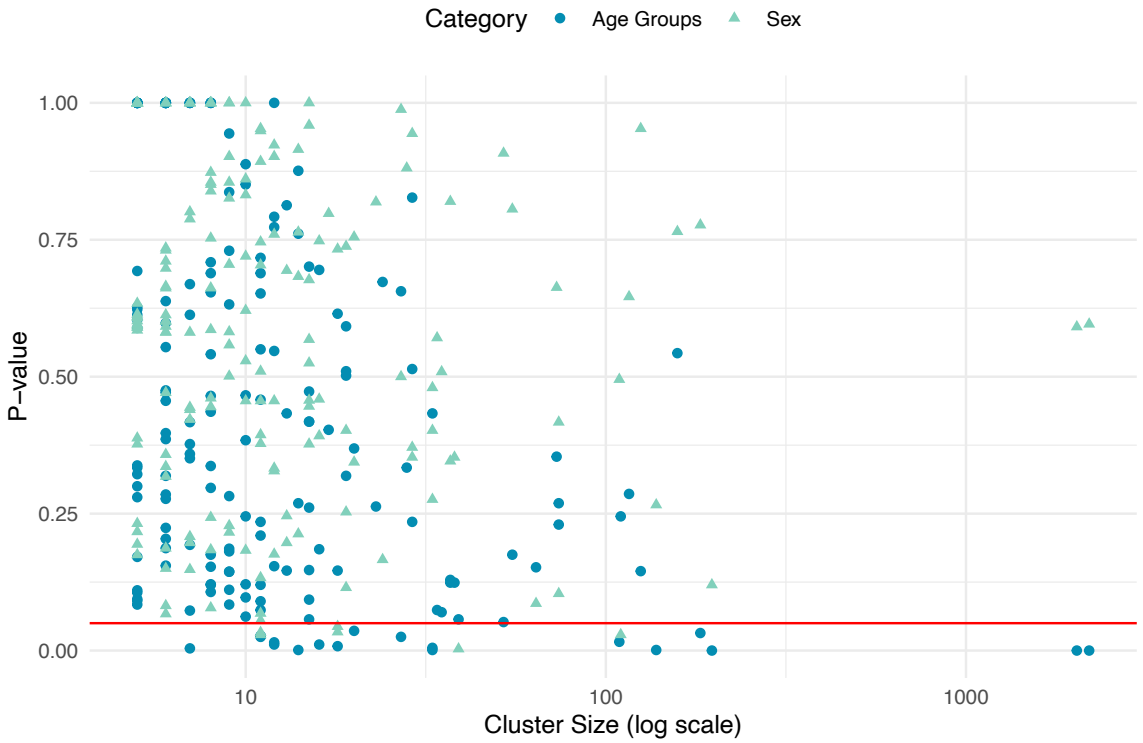

Figure 3. Phylogeny-Trait Correlation Among 181 Locally Circulating Clusters. Blue circles showed the test result of association between locally circulating clusters and age groups. Green triangles showed the test result against sex groups. Shapes bellowed the red line ($p < 0.05$) indicated clusters that has tightly correlated with corresponding traits.

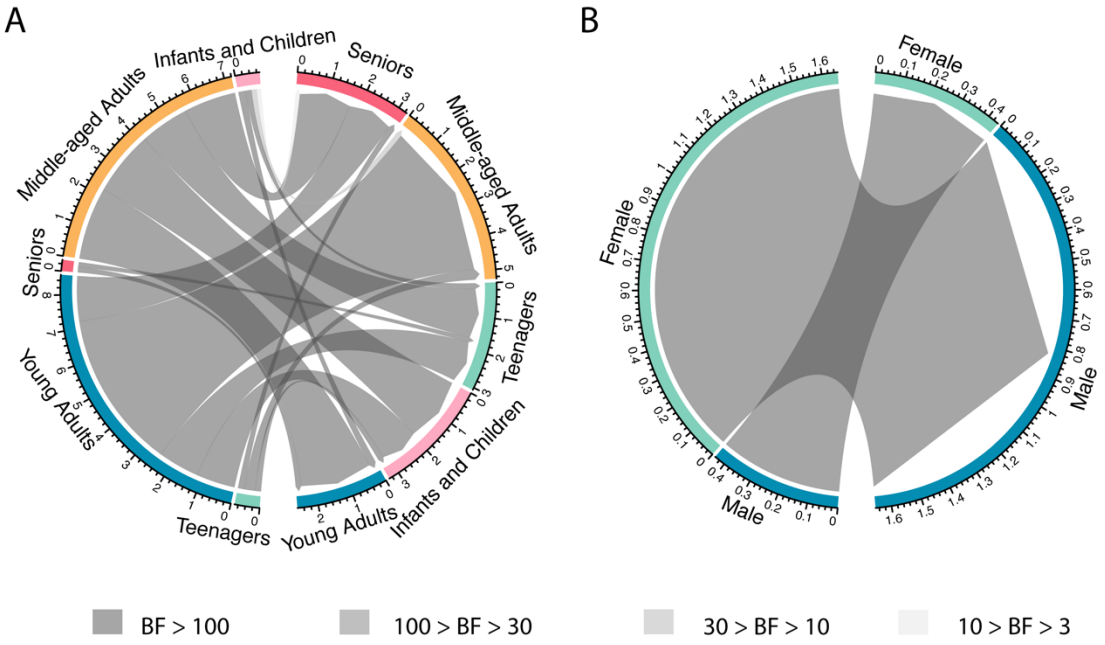

Figure 4. Discrete Trait Diffusion Models of Age Group and Sex. The age group model (**A**) and the sex model (**B**) are presented by circular charts. Chord thickness indicates the magnitude of the transition rate, while color signifies Bayes Factor support. Only transition rates supported by a Bayes Factor greater than 3 (BF > 3) from the discrete trait analysis are displayed. Transmission sources are on the left, transmission sinks are on the right.

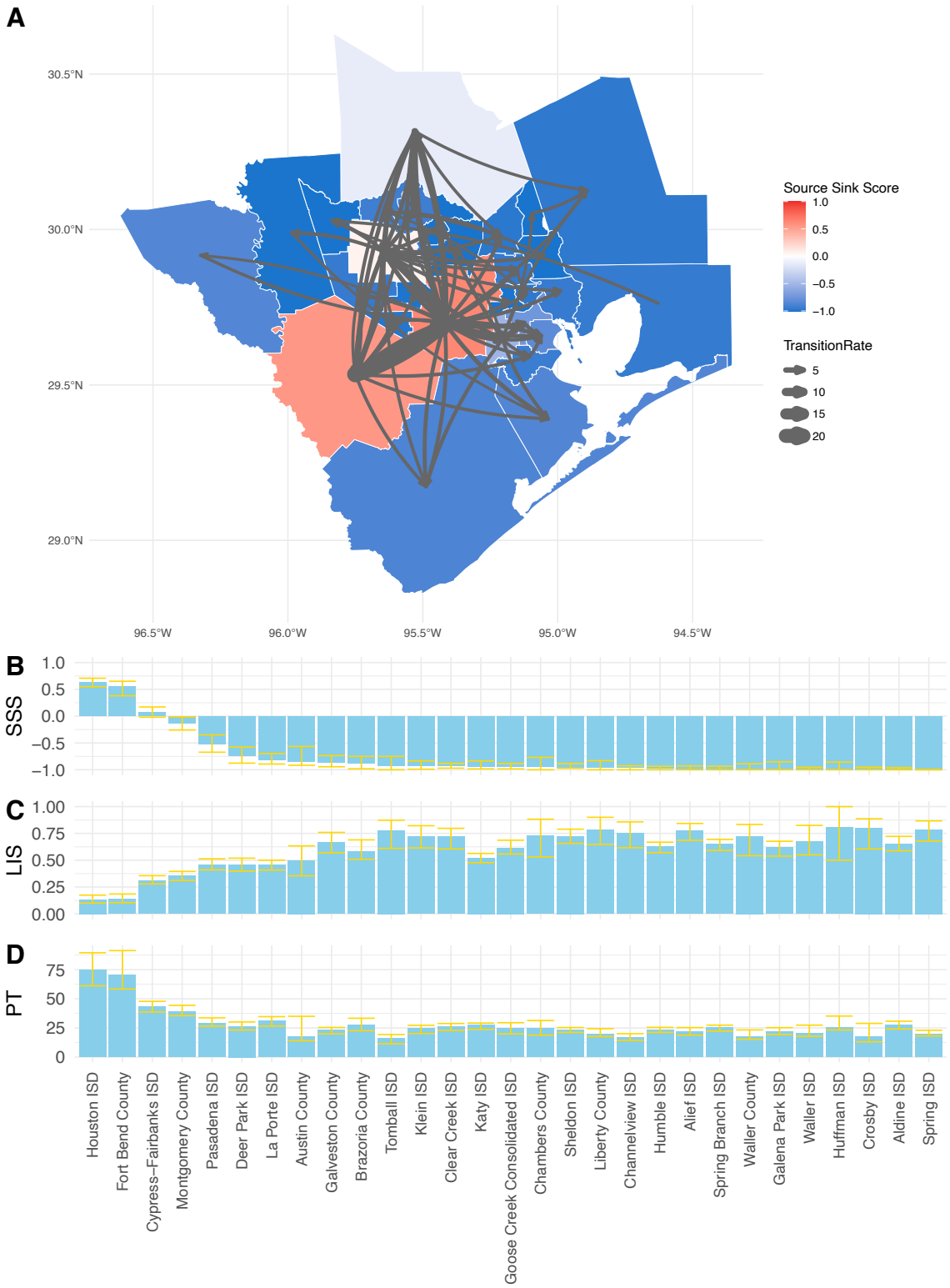

Figure 5. Distinct Transmission Patterns in Subregions of Greater Houston. (A) Discrete phylogeographic reconstruction of the dispersal history across 21 independent school districts in Harris County and 8 nearby counties. Subregions on the map are colored based on their associated

Source Sink Scores. Arrow thickness indicates the magnitude of transition rates. All transitions shown on the map were decisively supported by Bayes factors (>100) **(B)** Source Sink Scores (SSS), **(C)** Local Import Scores (LIS), and **(D)** Persistence Time (PT) across these subareas. In these bar charts, golden error bars represent the associated 95% Highest Posterior Density (HPD), providing a measure of uncertainty for each score. SSS ranges from 1 (viral source) to -1 (viral sink). LIS ranges from 0 (epidemic is locally maintained) to 1 (epidemic relies on introduction).